Comparing Fixed Effects and Covariance Structure Estimators

Ejrnæs, Mette; Holm, Anders

Publication date:
2004

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Ejrnæs, M., & Holm, A. (2004). *Comparing Fixed Effects and Covariance Structure Estimators*. Department of Economics, University of Copenhagen.

CAM

Centre for Applied Microeconometrics

Institute of Economics
University of Copenhagen

http://www.econ.ku.dk/CAM/

Comparing Fixed Effects and Covariance Structure Estimators

Mette Ejrnæs
Anders Holm

2004-02

The activities of CAM are financed by a grant from
The Danish National Research Foundation

Comparing Fixed Effects and Covariance Structure Estimators

Mette Ejrnæs*
Anders Holm†

January 7, 2004

Abstract

In this paper we compare the traditional econometric fixed effect/first difference estimator with the maximum likelihood estimator implied by covariance structure models for panel data. Our findings are that the maximum likelihood estimator is remarkable robust to mis-specifications, however in general the fixed estimator is preferable in small samples. Furthermore, we argue that we can use the Hausman test as a test of consistency of the maximum likelihood estimator. Finally we show that the covariance structure models is not identified in the case of time-invariant independent variables.

*CAM, Institute of Economics, University of Copenhagen.
†CAM and Departement of Sociology, University of Copenhagen.

1 Introduction

When analyzing panel data using linear regression models one faces the problem of clustered observations. However, panel data also offer a way to study the correlation between observed and unobserved variables. Ignoring clustering yields biased estimates of the variance of any estimated coefficients, see e.g. Diggle et al. (1994). Ignoring correlations yields biased estimates, see e.g. Wooldridge (2002). For example, if we study the effects of educational attainment on individual earnings in a simple linear model, we might find a significant positive least squares estimate of the effect of years of education on earnings. However, the problem is that we do not know if this is the "true" effect". The education variable might both pick up an genuine effect from education, but also effects from other variables not included in the model, such as different family background variables. If the influence from family background cannot be completely measured and thus taken into account in the model, the ordinary least squares (OLS) estimate of the effect of years of education will be biased. Hence additional modelling is required.

Two estimators have been proposed to take this into account. The fixed effect (FE) estimator, which only look at differences of earnings from unit means and Covariance structure estimators (CSM), that specifies a complete model for both observed and unobserved effects. The last estimator has just recently been proposed for panel data model in a paper by Teachman, Ducan, Yeung and Levy (2001). Both estimators take into account the clustered nature of panel data and also addresses the problem of unobserved variables being correlated with observed variables. Hence both estimators yield consistent estimators of the effect of the observed explanatory variables. But the two models differ in the complexity of underlying assumptions. On one hand, the fixed effect estimator is the one that requires the fewest assumptions, but on the other hand, the covariance structure estimator offers insight into the correlation structure between the observed and unobserved components of the model. None of the models allow the estimation of time invariant observed variables. That the fixed effect estimator only allow the estimation of time varying effects has been known for a long time, see Wooldridge (2002). In this paper we show that this also is the case for the CSM estimator.

The CSM estimator provides more insight about the correlation between the unobserved component and observed variables, and for this reason, it should be preferable. But the CSM estimator also relies on more assumptions, which may not be fulfilled. This suggest that one should apply the CSM

estimator but only when it is valid. Therefore, we propose a test which makes us able to tell when the more complicated CSM model offers consistent estimates from the perspective of the more simple fixed effect model.

The remaining paper is organized as follows. Section 2 presents the CSM and Fixed effect models, section 3 illustrates the performance of the estimators by a simulation study, section 4 discusses a test for comparing the fixed effect and the CSM estimator and section 5 has a small application of the estimators and finally section 6 offers some concluding remarks.

2 The model

In this section we will state a couple of different models in which we compare the covariance structure model (CSM) approach with the traditional panel data estimators: fixed effect (FE), First difference (FD) and random effect (RE). We will here discuss the properties of the different estimators in an analytical framework. In the next section the we further investigate the estimators in the different models by using simulation studies. As a baseline model we use the model stated in Teachman et al. (2001). This model is a panel data model including a fixed effect which potentially may be correlated with the explanatory variables. We will show that this model is rather restrictive and consider four different extensions of the model.

2.1 Baseline model

The baseline model is a standard panel data model where we allow for an individual unobserved effect. The key feature of this model is that all the explanatory variables are time varying and the errors are assumed normal. For simplicity we assume only one explanatory variable and that all variables are centered. None of the results rely on these assumptions. The model is given by

$$y_{it} = \beta x_{it} + \alpha_i + \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T$$

where x_{it} is the time varying explanatory variable, α_i is the individual specific effect, which is assumed to be normal $\alpha_i \sim N(0, \sigma_\alpha^2)$ and independent across individuals. In this setup we allow that this individual effect may be correlated with the explanatory variable such that $Cov(x_{it}, a_i) = \tau$ and $Cov(x_{it}, a_j) = 0$ if $j \neq i$. The last term ε_{it} is an idiosyncratic error term

which is assumed to be independent of all the other terms and is normally distributed $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$. The parameters are β, $\sigma_\alpha^2, \sigma_\varepsilon^2$ and τ, however our parameter of primarily interest is β.

As already pointed out in Teachman et al. (2001), the Random effect estimator of β is biased and inconsistent in this model unless $\tau = 0$. The two other estimators CSM and Fixed effect yield consistent estimates of β. The last estimator we focus on is the first difference estimator and it turns out that this estimator is also consistent.

The CSM estimator is based on following covariance structure

$$\begin{aligned} var(y_{it}) &= \beta^2 var(x_{it}) + 2\beta\tau + \sigma_\alpha^2 + \sigma_\varepsilon^2 && (1)\\ cov(y_{it}, y_{is}) &= \beta^2 cov(x_{it}, x_{is}) + 2\beta\tau + \sigma_\alpha^2 \quad s \neq t && (2)\\ cov(y_{it}, x_{it}) &= \beta var(x_{it}) + \tau && (3)\\ cov(y_{it}, x_{is}) &= cov(y_{is}, x_{it}) = \beta cov(x_{it}, x_{is}) + \tau \quad s \neq t && (4) \end{aligned}$$

Since we assume independence across individuals, all covariances across individuals zero. The four parameters of the model is identified from the covariance structure and the estimation is performed by applying the maximum likelihood approach. The likelihood function for the model is given by:

$$\ln L = \ln |\Sigma(\theta)| + tr\left(S\Sigma^{-1}(\theta)\right)$$

expect from an additive constants and where $\theta = (\beta, \tau, \sigma_\alpha^2, \sigma_\varepsilon^2)$, S is the sample covariance matrix of (y, x) and $\Sigma(\theta)$ is the CSM covariance structure given by equation (1)-(4).

The first differences and the fixed effect estimator are obtained by transforming the model such that the individual effect cancels out. The first difference estimator is based on first differences such that the model is transform

$$\begin{aligned} y_{it} - y_{it-1} &= \beta x_{it} + \alpha_i + \varepsilon_{it} - \beta x_{it-1} - \alpha_i - \varepsilon_{it-1}, \qquad i = 1, \ldots, N, \quad t = 2, \ldots, T\\ \Delta y_{it} &= \beta \Delta x_{it} + \Delta \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 2, \ldots, T. \end{aligned}$$

The Fixed effect estimator (or Within estimator, which the fixed effect estimator is sometimes also called) can be obtained from the following transformation

$$\begin{aligned} y_{it} - y_{i.} &= \beta x_{it} + \alpha_i + \varepsilon_{it} - \beta x_{i.} - \alpha_i - \varepsilon_{i.}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T\\ y_{it}^* &= \beta x_{it}^* + \varepsilon_{it}^*, \qquad i = 1, \ldots, N, \quad t = 2, \ldots, T, \end{aligned}$$

where $y_{i.} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $x_{i.} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$ and $\varepsilon_{i.} = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}$.For both estimators the estimate of β is obtained by running an OLS regression on the transformed model. The estimators are given by

$$\hat{\beta}_{FD} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T} \Delta y_{it}\Delta x_{it}}{\sum_{i=1}^{N}\sum_{t=2}^{T} \Delta x_{it}^2}, \hat{\beta}_{FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}^* x_{it}^*}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^{*2}}$$

In the case where $T = 2$ the FD and FE are identical.

To establish the link between the CSM estimator and the First Difference estimator, we can show that FD and FE is also based on the covariance structure outlined for the CSM approach. However, instead of using all four moment conditions to estimate parameters, the FD and FE estimator only uses a linear combination of the moment conditions (In the appendix the it is shown how one can derive the FD on the basis of the moment conditions).

In the baseline model the CSM, FE and FD are consistent, but CSM is asymptotically more efficient than the other two because it is equivalent to MLE. The RE estimator is inconsistent. Therefore, the CSM model is preferable in large samples. However, what we will show is that the CSM estimator is less robust when relaxing some of the assumption in the model. In the following we relax the model in three different ways.

2.2 A Model with heterogeneity

In this model we allow for heterogeneity in the correlation between the individual effect and the explanatory variables. This can be done by assuming that $Cov(x_{it}, \alpha_i) = \tau_i$. What we assume here is that some individuals may have a positive correlation while other can have a negative correlation, and that the size of the correlations might differ in size.

In this case the CSM is misspecified.[1] In the simulation studies in the next section we show how this affects the CSM estimator. On the other hand one can easily show that this extension does not affect the FD or FE estimators. For both estimators the transformation of data is still valid.

[1] One can of cause introduce these N new parameters $\tau_1, .., \tau_N$ and the model is no longer misspecified. However, the purpose of this paper is to investigate the CSM estimator to the baseline model. Furthermore, by allowing individual correlations, we no longer have large N asymptotics for the CSM estimator, see e.g. Night (2000).

2.3 A Model with Non-normal errors

In the baseline model we have assumed normality for all the error terms and the individual effect. This assumption may be rather restrictive, since in a lot of contexts we do not know the distribution of the errors. In particular, one can think of examples where the distribution of the individual effect is non-normal. As an extreme, one can assume that the individual effect can take two values. In this case the distribution is a binomial distribution and is characterized by two parameters: the probability p and the mass point $w : \Pr(a_i = w) = p$. The other mass point is given by $\frac{-wp}{(1-p)}$ and the probability is $1 - p$.[2]

However, for the FD and FE estimator normality is not necessary, while the CSM estimator relies on normality. What we will need in order to obtain consistent an estimate of β using either FD or FE is that the errors have mean zero and a finite variance. The reason why FD and FE does not rely on normality is because they are general moment estimators, see e.g. Woolridge (2002). The impact of non normal individual effects for the CSM is examined in the simulation studies. We also study the effect of non-normality of the idiosyncratic error term.

2.4 A Model with Non-linear dependency between α and x

In the baseline model it is assumed that the explanatory variable x and the individual effect α are joint normally distributed. This implies that x can be written as a linear function of α.However, one can also imagine cases where x is a non-linear function of α.If x is a non-linear function of the individual effect

$$x_{it} = f(\alpha_i)$$

the assumption of joint normality is violated. A particular example is when the individual specific effect does not only influence the conditional mean of the variable x but also the conditional variance. Let the explanatory variable be given by

$$x_{it} = (\alpha_i^2) * (\xi_{it} + \frac{\tau \alpha_i}{3}) \qquad \xi_{it} \sim iiN(\mu_x, \sigma_\xi^2),$$

[2]This condition insures that the mean of the individual effect is zero.

where ξ and α are mutual independent.[3] In this example the conditional mean and variance of x is given by

$$\begin{aligned} E(x_{it}|\alpha_i) &= (\alpha_i^2) * (\mu_x + \frac{\tau\alpha_i}{3}) \\ V(x_{it}|\alpha_i) &= \alpha_i^4 * \sigma_\xi^2. \end{aligned}$$

In this case the CSM model is misspecified, while FD or FE are still consistent because they do not rely on any distributional assumption of the explanatory variables and the unobserved effect. The fact, that for both FD and FE , the individual effect is eliminated by a transformation, means that it does not cause any problem that the explanatory variable is any function of the uobserved individual effect, α.

2.5 A Model with a time invariant variable

In the baseline model, we only considered time varying explanatory variables. However, often in empirical analyses, one also would like to include time invariant variables. The model with time invariant variables is given by

$$y_{it} = \beta x_{it} + \gamma z_i + \alpha_i + \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where z_i is the time invariant explanatory variable. We also allow that z_i can be correlated with α_i such that $cov(\alpha_i, z_i) = \rho$.For the time varying explanatory variable and the error terms the assumptions are as in the baseline model. In this model we have six parameters: $\beta, \gamma, \sigma_a^2, \sigma_e^2, \tau$ and ρ.

In this model the covariance structure is given by:

$$\begin{aligned} var(y_{it}) &= \beta^2 var(x_{it}) + \gamma^2 var(z_i) + 2\gamma\beta cov(x_{it}, z_i) + 2\beta\tau + 2\gamma\rho + \sigma_\alpha^2 + \sigma_\varepsilon^2 \\ cov(y_{it}, y_{is}) &= \beta^2 cov(x_{it}, x_{is}) + \gamma^2 var(z_i) + 2\gamma\beta cov(x_{it}, z_i) + 2\beta\tau + 2\gamma\rho + \sigma_\alpha^2 \quad s \neq t \\ cov(y_{it}, x_{it}) &= \beta var(x_{it}) + \gamma cov(x_{it}, z_i) + \tau \\ cov(y_{it}, x_{is}) &= cov(y_{is}, x_{it}) = \beta cov(x_{it}, x_{is}) + \gamma cov(x_{is,} z_i) + \tau \quad s \neq t \\ cov(y_{it}, z_i) &= \beta cov(x_{it}, z_i) + \gamma var(z_i) + \rho. \end{aligned}$$

In this model the parameters in is not identified with a CSM estimation. The easiest way of verifying this, is by seeing that we introduce two new parameters to the model: γ and ρ, but the new model only give rise to one

[3] The functional form is chosen such that $cov(\alpha_i, x_i) = \tau$.

extra equation in the covariance matrix namely $cov(y_{it}, z_i)$.The difference between introducing a time invariant variable and a time varying variable w is that the time varying variable generates two extra equations in the covariance structure: $cov(y_{it}, w_{it})$ and $cov(y_{is}, w_{it})$.

However, in contrast to the CSM model which is not identified the FD and FE can still obtain consistent estimate of β. The reason why FD and FE are still valid is because the transformation are still valid. By performing the transformation outlined in the baseline model, the time invariant variable cancels out. This means that one is not able to identify the effect of the time invariant variable but one can get a consistent estimate of β. The conclusions about the estimators hold irrespective of whether τ and ρ are zero. One can of course also estimate the CSM model while absorbing the time independent variable into the uobservable individual effect, α.

In contrast to the CSM, FD and FE the RE estimator can produce an estimate of γ. However, this estimate is only consistent if $\tau = \rho = 0$.

3 Simulation studies

In the previous section we outlined various extensions of the baseline model, in which the CSM estimator was misspecified. In order to investigate the impact of the misspecification, we conduct a number of simulation studies where we compare the different estimators. In all the models we consider the two cases: no correlation and a positive correlation between individual effects and the explanatory variable. In the simulation study, we report the average and standard error for the different estimators. Furthermore, we also perform a test for no correlation between the individual effect and explanatory variable. For the CSM estimation this test can be formulated as a hypothesis directly on the estimated parameter of the covariance τ :

$$H_0 : \tau = 0.$$

The test in the simulation study is conducted as a Wald test, where the test statistics is given by

$$q = \frac{\hat{\tau}^2}{var(\hat{\tau})} \overset{a}{\sim} \chi^2_{(1)}.$$

The usual test for correlation between individual effect and explanatory variables is performed using a Hausman test, Hausman (1978). The underlying

idea of this test is that if there is no correlation between the individual effect and the explanatory variable the both the RE estimator and the FE (or FD) estimator are consistent, but only the RE estimator is efficient. If there is a correlation then only FE (or FD) are consistent. This implies that if there is no correlation the RE estimator and the FE estimator are close. The Hausman test is not a direct test on the covariance but and indirect test. The test and the test statistics are formulated as

$$\begin{aligned} H_0 &: \quad \text{No correlation between } \alpha_i \text{ and } x_{it} \\ H &= \frac{\left(\hat{\beta}_{RE} - \hat{\beta}_{FE}\right)^2}{V(\hat{\beta}_{FE}) - V(\hat{\beta}_{RE})} \end{aligned}$$

where $V(\hat{\beta}_{FE})$ and $V(\hat{\beta}_{RE})$ are the variances of $\hat{\beta}_{RE}$ and $\hat{\beta}_{FE}$ respectively.

The simulation experiment is set up such that each simulation consists of 250 units observed over two time-periods, which makes 500 observations. Total number of replications are 10,000. In all the simulations, the variances of α_i and ε_{is} are set to one.

In table 1, the baseline model has been simulated for two values of τ. The first two columns refers to the case where $\tau = 0$ indicating no correlation between α_i and x_{it}. As expected, all the considered estimates of β are centered around the true value. Moving to the case where $\tau = 0.25$ we find that both OLS and RE-effect estimator is biased while CSM and FE remain unbiased. The bias of the OLS estimator is of the magnitude 12 percent and the bias of RE estimator of about 6 percent in this simulation. The CSM estimates of τ are in both cases unbiased.

The variance of the estimators of β in this estimation study suggests that the variance of the FE and the CSM estimates are of the same magnitude, while the RE-estimates have lower variance.

Turning to the test for no correlation between α_i and x_{it} we find that the Hausman test reject (at a 5 percent significance level) the true hypothesis in 5.0 percent of the cases where the CSM test rejects the true hypothesis in 17.3 percent of the cases. Hence, this simulation example actually indicates that the Hausman test may be preferable to the CSM test, since it rejects "too often" a true hypothesis. When the hypothesis is not true the CSM test seem to have more power in rejecting a wrong hypothesis; CSM test rejects in 91 percent of the cases where the Hausman test only rejects in 76 percent of the cases.

Table 1. Baseline model.

	$\tau = 0$		$\tau = 0.25$	
Estimator	Mean	std	Mean	std
OLS β	1.0004	0.0451	1.1213	0.0447
Fixed Effect β	1.0004	0.0446	1.0004	0.0446
Random Effect β	1.0004	0.0388	1.0616	0.0392
CSM β	0.9958	0.0447	0.9962	0.0447
τ	0.0043	0.0904	0.2533	0.0955
Fraction of rejections of $H_0 : \tau = 0$ at a 5 percent significance level				
CSM model (t-test)	0.1733		0.9124	
Hausman-test (FE against RE)	0.0503		0.7648	

Note: $\beta = 1, \alpha_i \sim iiN(0,1), \varepsilon_{it} \sim iiN(0,1)$,
$x_{it} = \xi_{it} + \tau\alpha_i, \xi_{it} \sim iiN(0,2)$

In table 2, we simulated the model with heterogenous correlation between α_i and x_{it} described in section 2.2. In this particular study we assume that for 50 percent of the individuals there are no correlation between α_i and x_{it}, while for the remaining 50 percent they have a correlation corresponding to $\tau = 0.5$. The simulation results seem almost unaffected by the fact that the correlation is heterogenous. The same seem to be the case when change the distribution of the individual effect from a normal distribution to a binomial distribution (see table 3). This suggests that the CSM model is very robust to these two types of misspecification.

Table 2. Heterogenous correlation.

	$\tau = 0$		$\bar{\tau} = 0.25$	
Estimator	Mean	std	Mean	std
OLS β	1.0004	0.0451	1.1185	0.0449
Fixed Effect β	1.0004	0.0446	1.0004	0.0446
Random Effect β	1.0004	0.0388	1.0611	0.0391
CSM β	0.9958	0.0447	0.9960	0.0447
τ	0.0043	0.0904	0.2557	0.1002
Fraction of rejections of $H_0 : \tau = 0$				
CSM model (t-test)	0.1733		0.8967	
Hausman-test (FE against RE)	0.0503		0.7396	

Note: $\beta = 1, \alpha_i \sim iiN(0,1), \varepsilon_{it} \sim iiN(0,1), x_{it} = \xi_{it} + \tau_i\alpha_i$,
$\xi_{it} \sim N(0,2), Pr(\tau_i = 0) = \Pr(\tau_i = 0.5) = 0.5$

Table 3. Non-normal individual specific effect

	$\tau = 0$		$\tau = 0.25$	
Estimator	Mean	std	Mean	std
OLS β	1.0002	0.0451	1.1213	0.0443
Fixed Effect β	1.0003	0.0450	1.0003	0.0450
Random Effect β	1.0003	0.0391	1.0616	0.0393
CSM β	0.9955	0.0450	0.9959	0.0449
τ	0.0044	0.0893	0.2534	0.0933
Fraction of rejections of $H_0 : \tau = 0$				
CSM model (t-test)	0.1692		0.9185	
Hausman-test (FE against RE)	0.0467		0.7720	

Note: $\beta = 1, \varepsilon_{it} \sim iiN(0,1), x_{it} = \xi_{it} + \tau\alpha_i, \xi_{it} \sim N(0,2)$
$\Pr(\alpha_i = \sqrt{3}) = 0.25, \Pr(\alpha_i = -\frac{1}{\sqrt{3}}) = 0.75$

In table 4, the model has been changed such that there is a non-linear relation between α_i and x_{it}.In the case with no correlation between α_i and x_{it} we have that OLS, FE and RE are unbiased, but the CSM estimator of β is biased. The magnitude of the bias is about 4 percent. Furthermore, the CSM estimate of τ is also biased. When we move to the case with a correlation between between α_i and x_{it} we still get that OLS and RE is biased but also the CSM. The magnitude of the bias of the CSM model is at the same level as the the RE model. This indicates that in cases with non-linear relationship between α_i and x_{it} the CSM estimator do not perform well.

Table 4. Non-linear relation between α and x

	$\tau = 0$		$\tau = 0.25$	
Estimator	Mean	std	Mean	std
OLS β	1.0007	0.0636	1.0809	0.0601
Fixed Effect β	0.9995	0.0387	0.9995	0.0387
Random Effect β	1.0001	0.0426	1.0418	0.0423
CSM β	0.9668	0.0513	0.9696	0.0489
τ	0.0545	0.1967	0.2944	0.2191
Fraction of rejections of $H_0 : \tau = 0$				
CSM model (t-test)	0.4399		0.7686	
Hausman-test (FE against RE)	0.2487		0.5741	

Note: $\beta = 1, \varepsilon_{it} \sim iiN(0,1), x_{it} = \alpha_i^2(\xi_{it} + \frac{\tau\alpha_i}{3})$,
$\alpha_i \sim iiN(0,1), \xi_{it} \sim N(0,2)$

To sum up the results of the four simulation study, we find that the FE estimator is superior to the CSM estimator because it does not exhibit any

bias in then four different set ups and variance of the estimator is almost identical to the CSM estimator. Furthermore, when testing for correlation between the individual effect and the explanatory variable the CSM estimator also seem to have a problems by rejecting the true hypothesis too often. However, the simulation study also suggested that CSM estimator is robust to misspecifications with respect to the distributional assumptions and heterogeneity in the correlation. Moreover, in the cases when the CSM provides an unbiased estimate of β it also seems to yield unbiased estimate of τ.

4 A test of consistency of the CSM model

In the previous section we illustrated by a simulation study that in some cases of misspecification the CSM estimator do not perform well. This suggests that one should use the FD or FE estimator instead, which are very robust to misspecifications. However, one of the shortcomings of the FE and FD is that they does not provide any information about the correlation between the individual effect and the explanatory variable, except that one can test whether the correlation is zero. Therefore, one may prefer to use the CSM estimator, when it is consistent. To find out if the CSM estimator is consistent, we suggest to apply a Hausman test.

The idea is to compare the CSM estimator with the FE estimator which are consistent under weak assumptions. If there is a large difference between the estimators, this suggest that the CSM estimator is not consistent. In order to derive the test, we assume that the FD or FE estimators are consistent. We can then employ the Hausman-test as a consistency test of the CSM estimator of β, see Lee (1996) for some other general applications of the Hausman test:

$$\begin{aligned} H_0 &: \quad \hat{\beta}_{CSM} - \hat{\beta}_{FE} = 0 \\ H &= \frac{\left(\hat{\beta}_{CSM} - \hat{\beta}_{FE}\right)^2}{V(\hat{\beta}_{FE}) - V(\hat{\beta}_{CSM})} \sim \chi^2(1). \end{aligned}$$

In case this test is accepted there is also reason to believe that τ_{CSM} is consistent. Hence the Hausman test can be used as a specification test of the CSM model, and we can use the CSM model to obtain consistent information on τ.

5 A case study

As an illustration of the previous discussion, we will apply the FE and the CSM estimator to the case of earnings among siblings. The data are taken from the Panel Study of Income Dynamics (PSID) like in Teachman et al. (2001). We study the earnings from siblings from a random selection of families from wave 1988 to 1992. We use information on earnings and years of education when each of the siblings are 36 during the sample period. By making this restriction we do not have to take any age dependency into consideration in our modelling. We also restrict the analysis to individuals with none-zero earnings. All earnings are discounted to 1992 price level. Both types of restrictions means that some individuals appear in the data without siblings even though they actual might have some.

We propose a model where individual earnings depends on some individual characteristics: gender, years of education and birth order and some family characteristics, fathers education. We also assume that there may be unobserved family specific effects, such as the family genome and parental behavior that affect earnings of the children. As both behavior as well as genetics[4] are very likely both to affect the children's earnings capacity over and above that induced by their educational choice, as well as their educational choice, it is equaly likely that the family specific effect will be correlated with the educational attainment of the children. Hence, we expect that any estimator (OLS, RE) that does not take this into account will be biased. Formerly our model is:

$$y_{it} = \beta_0 + \Sigma_j \beta_j x_{jit} + \alpha_i + \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T$$

where i now indexes family units and t children. In this model the term α_i captures the family specific effect.

In table 5 we show some summary statistics of the data.

[4]Individuals with non-biological parents (i.e. siblings with fathers with different education) has been eliminated from our sample.

Table 5. Summary statistics

Variable	Mean	sd.dev	min	max
ln(Income)	2.304	0.711	-1.061	4.723
Years of education	13.366	2.057	6	17
Years of education (father)	12.794	1.916	6	14
Women	0.531	-	0	1
no. of children per family	1.404	-	1	8
No. of individuals	868			
No. of individuals with siblings	442			
No. of families	614			

The data is unbalanced in the sense that the different families in the might have different number of children, and some families appears to have only one child.

For comparison we estimate the model by using four different approaches: OLS, RE, FE and CSM. When estimating the model with OLS, RE and parts of the CSM, that measures inter family correlations we can include the variables that are the same for all children within a family such as father's education, but when estimating the model with FE or the part of the CSM where measures of intra family correlations, these have to be omitted. This means that the family specific effect in the FE and CSM model also contains the impact of father's education. The covariance structure of the model is shown below. It only differs from the structure shown in 1-4 due to the inclusion of more explanatory variables:

$$var(y_{it}) = \Sigma_l \Sigma_j \beta_j \beta_l cov(x_{jit}, x_{lit}) + 2\Sigma_j \beta_j \tau_j + \sigma_\alpha^2 + \sigma_\varepsilon^2 \tag{5}$$

$$cov(y_{it}, y_{is}) = \Sigma_l \Sigma_j \beta_j \beta_l cov(x_{jit}, x_{lis}) + 2\Sigma_j \beta_j \tau_j + \sigma_\alpha^2, \quad s \neq t \tag{6}$$

$$cov(y_{it}, x_{jit}) = \Sigma_l \beta_l cov(x_{jit}, x_{lit}) + \Sigma_l \tau_l, \quad j = 1, 2 \tag{7}$$

$$cov(y_{it}, x_{jis}) = \Sigma_l \beta_l cov(x_{jit}, x_{lis}) + \Sigma_l \tau_l, \quad j = 1, 2, \quad s \neq t \tag{8}$$

Estimation results for different estimators of the model is shown in table 6 below.

Table 6. Estimation results.

Variable	OLS	RE	FE	CSM
Constant	0.779	0.931	−	−
	(.146)	(0.141)	(−)	(−)
Years of education	0.111	0.099	0.062	0.055
	(0.012)	(0.012)	(0.022)	(0.013)
Father's education	0.0390	0.048	−	−
	(0.013)	(0.015)	(−)	(−)
Gender	−0.291	−0.331	−0.418	−0.339
	(0.044)	(0.037)	(0.061)	(0.038)
τ_1	-	-	-	0.317
				(0.048)
τ_2	-	-	-	0.011
				(0.006)
σ_α^2	-	0.209	-	0.160
σ_ε^2	0.411	0.202	0.309	0.211

First, we test if the family specific effect is correlated with the explanatory variables. This is done by a Hausman test where the RE estimates are compared the FE estimates. The Hausman test statistic for the consistency of the RE model against the FE model is $7.12 \sim \chi^2(2)$. Thus the RE model is rejected against the FE model, which indicates that the family specific effect is correlated with the time varying explanatory variables[5]. This means that OLS and RE will produce biased estimates. When comparing the OLS and RE estimates we find a much higher return to education than for the FE and CSM model, which suggests that OLS and RE overestimates the return to education (as expected).

As shown in the previous section both the FE and CSM estimates are consistent when there is a correlation between the family specific effect and explanatory variables, but the FE is consistent under weaker assumptions. To test if CSM estimates are consistent we perform the Hausman test described in section 4. The value of the Hausman test statistic for consistency of the CSM model (CSM against FE) is: $4.77 \sim \chi^2(2)$ which is insignificant (at a 5 % level). Hence, the CSM model is consistent according to the test and the average correlations between the family specific effects and the

[5]Note the the RE model takes into account farthers education. Hence, other unobserved unit invariant variables must generate the correlation between the time varying observed variables and the unobserved familiy variables.

observed explanatory variables are consistently estimated. The estimates of the correlation indicate that the family specific effect is positively and significantly correlated with years of education, while we find an insignificant correlation with the dummy for women. One interpretation of this result is that on one hand a "good" family background increases both the average earning capacity (above the level imposed by the educational attainment) and average educational level of the children from the family. On the other hand a good family background is not affecting the gender mix within a family. (fraction of daughters in a family). These results seem plausible and in accordance of what we expected. Finally, in this model we can perform a test of hypothesis that family specific effects are uncorrelated with the explanatory variables. The hypothesis is given by $H_0 : \tau_1 = \tau_2 = 0$. The Wald test-statistic is $12.45 \sim \chi^2(2)$, also rejects the hypothesis that the family effect is uncorrelated with the explanatory variables. The CSM estimates indicate that the annual earnings increases by about 5.5 percent per extra year of education.[6] Furthermore we find that women earn about 34 percent less than men.

To summarize the results, we find evidence for the existence of a correlated family specific effect in the earnings equation. This implies that the coefficient for years of education is biased for the OLS and RE. When performing a Hausman test comparing the FE and the CSM estimates, we find that the CSM estimates are valid and we can use the estimates of the covariances between the family effect and explanatory variables. We find that years of education are positively correlated with the unobserved family effect, and for this reason the pay-off from years on education are exaggerated for the OLS and RE estimators.

6 Conclusion

In this paper we have discussed modelling panel data with fixed effects. We have shown, as many others have before us, that the Ordinary least squares estimator and the random effects estimator yields biased estimates when time invariant unobserved effects are correlated with the observed explanatory variables. Two alternative estimators have been proposed in the literature, namely the covariance structure estimator and the fixed effect estimator.

[6] Although we have controlled for family specific unobsered effects, one may still consider if this esimate is a true estimate of the return to education.

Both take into account the possible correlation between observed and unobserved variables. However, they differ in the how they approach the problem. The Fixed effect estimator use a concentrated likelihood approaches and sweeps out the fixed effect, whereas the covariance structure estimator specifies the correlation between the unobserved and observed variables. This leaves the latter approach more vulnerable to misspecification but also offers more insight into the structure of the data. In this paper we examines in details how the misspecifications affect the covariance structure estimator and suggest a way to test whether it is mispeccified against the more robust fixed effect estimator in actual applications.

7 Appendix. Deriving the First difference estimator from the covariance equations.

The FD estimator is based on the difference between $cov(y_{it}, x_{it})$ and $cov(y_{it}, x_{is})$ when $s = t - 1$

$$\begin{aligned}
[cov(y_{it}, x_{it}) - cov(y_{it}, x_{it-1})] &= \beta(var(x_{it}) - cov(x_{it}, x_{it-1})) \Leftrightarrow \\
[E(y_{it}x_{it}) - E(y_{it}x_{it-1})] &= \beta(E(x_{it}^2) - E(x_{it}x_{it-1}) \Leftrightarrow \\
E(y_{it}\Delta x_{it}) &= \beta E(x_{it}\Delta x_{it}).
\end{aligned}$$

Similar one can get that

$$\begin{aligned}
[cov(y_{it-1}, x_{it}) - cov(y_{it}, x_{it-1})] &= \beta(cov(x_{it}x_{it-1}) - var(x_{it-1})) \Leftrightarrow \\
E(y_{it-1}\Delta x_{it}) &= \beta E(x_{it-1}\Delta x_{it}).
\end{aligned}$$

By subtracting these two expressions we obtain

$$\begin{aligned}
E(y_{it}\Delta x_{it}) - E(y_{it-1}\Delta x_{it}) &= \beta E(x_{it}\Delta x_{it}) - \beta E(x_{it-1}\Delta x_{it}) \Leftrightarrow \\
E(\Delta y_{it}\Delta x_{it}) &= \beta E(\Delta x_{it}^2) \Leftrightarrow \\
\beta &= \frac{E(\Delta y_{it}\Delta x_{it})}{E(\Delta x_{it}^2)}
\end{aligned}$$

By replacing the moment conditions with the sample moment condition one can obtain an estimate of β,which is equivalent to the FD estimator of β.

References

[1] Diggle, Peter, Kung-Yee Liang and Scott Zeger 1994. "Analysis of Longitudinal Data". Oxford University Press.

[2] Hausmann, James 1978. "Specification Tests in Econometrics." Econometrica 46: 69-85.

[3] Lee, M. (1996) "Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models", Springer.

[4] Night, K. (2000), Mathematical Statistics, Chapman & Hall/CRC.

[5] Teachman, Jay, Greg J. Ducan, W. Jean Yeung and Dan Levy 2001. "Covariance Structure Models for Fixed and Random Effects." Sociological methods and Research 30: 242-270.

[6] Wooldrigde, Jeffrey 2002. "Econometric Analysis of Cross Section and Panel Data". MIT press.